DOCUMENT RESUME

ED 305 367 TM 012 692

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE Differential Item Functioning, Item Discrepancy and Bias.
PUB DATE 88
NOTE 20p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Black Students; College Entrance Examinations; Higher Education; Hispanic Americans; *Item Analysis; Latent Trait Theory; Licensing Examinations (Professions); *Racial Differences; Scores; Statistical Analysis; *Test Bias; Test Items; White Students
IDENTIFIERS *Differential Item Performance; *Item Bias; Standardization

ABSTRACT

The standardization method of determining differential item functioning (DIF) is not acceptable for detecting biased items because of its methodological limitations. The main problems with the standardization index by N. J. Dorans and E. Kulick (1983) are: (1) it does not detect as DIF all items with large or aberrant group differences; (2) it considers as DIF items with average differences in passing the item equal to the average difference in ability between the two groups; and (3) in normal circumstances, any item will be considered as DIF if the difference in passing the item at each ability level is 8% beyond the difference in ability level of the groups measured by the percentage answering all items correctly. Applying this procedure to Scholastic Aptitude Test data for White, Black, and Hispanic students demonstrates its limitations. Several practical problems are also found when the procedure is applied to a licensure test. A practical method of detecting biased items is to use the differences in percentages passing the item among two groups. This simple method can reduce racial differences in test scores, as illustrated by its application to the results of an actual licensure examination that was administered to 343 applicants. Two tables present data on both techniques. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED305367

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Eulogio Gonzalez

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DIFFERENTIAL ITEM FUNCTIONING, ITEM DISCREPANCY AND BIAS

Eulogio Gonzalez-Tamayo

New York, 1988

RUNNING HEAD: Dif, Discrepancy and Bias

SUMMARY

The Standardization Procedure of determining differential item functioning is rejected as a method of detecting item biased. The procedure presents both technical and practical problems when applied to a licensure test. The differences in percentage passing the item among two groups is proposed as a practical method of detecting items bised representing a simple, verifiable, and practical step to reduce unnecessary racial differences in test scores.

The issue of test or item bias has attracted considerable attention during the last twenty years. The purpose of any psychological test is to differentiate among individuals and groups; a test that does not differentiate is useless. Test scores, on the other hand, are only indirect and imprecise indicators of the trait and they do not represent the real thing. Under these circunstances, the question arises about whether or not test scores can accurately differentiate among groups markedly differen in background and, assuming that they can, an additional problem is to determine how much of the found differences are real differences in the attribute and how much is due to artifacts of the imperfect techniques used to measure the trait.

The meaning of the word bias in statisitical terms is a constant and systematic error of measurement as opposed to chance error (Anastasi, 1982) or a systematic under- or overestimation of a population parameter by a statistic based on samples drawn from the population (Jensen, 1980, 1984; Reynolds & Brown, 1984). The word bias, however, has some loaded meaning and it represents an interpretation or evaluation of statistical facts. For this reason, recently, researchers have been using more neutral and descriptive words like discrepant, aberrant or anomalous items, differential items functioning (dif) and differential item performance (dip). Statistical techniques are used to detect anomalies in what was believed to be an homogeneous set of items. Whatever it is that the test is measuring, the biased item behaves differently and bias is detected when an item does not fit the pattern established by other items in the test. Statistical tecniques do not answer questions about the origin of the found group discrepancies, about whether or not the test should be used even if it is unbiased, or about how the found discrepancies should be interpreted.

The mere fact of group differences in mean scores is not a sufficient basis for a claim of discriminatory bias in an item or in a test since this position would eliminate the purpose of testing which is to determine differences in the trait. Mean differences, on the other hand, do not by themselves indicate differences in the trait Researchers have tried to separate psychometrically detectable bias in the test from possible differences on the trait but the efforts have not been completely satis-

actory (Ironson, 1982). It is, therefore, necessary that bias procedures (1) make adjustments for group differences in order to separate bias from real differences on the trait and (2) make provisions to determine in an objective way the group standin on the trait. The total score on the test may be the most accessible and the best available way of determining group and individual standing on the trait but it is not a completely defensible measure when the test is analyzed for possible bias. Operational definitions here and in any other place trade simplicity and clarity for accuracy (Scriven, 1980).

Several statistical and judgmental methods of detecting item bias have been proposed from time to time (Berk, 1982). Latent trait models are sometimes regarded as the standard by which all other methods are to be compared due to their theoretical sofistication. The analyses, however, are complex, expensive, and impractical in most licensing situations due to the requirements of unidimensionality, large samples of subjects and complex statistical calculations. Two recent methods of detecting differential item functioning proposed by researchers at ETS are the Standardization Method (Dorans & Kulick, 1986) and the Mantel-Haenszel (Holland & Thayer, 1988; Mantel-Haenszel, 1959) procedure. Both methods seems to be acceptable theoretically because they make provisions for initial determination of differences in group means and for the separation of bias and real differences in the trait. Both techniques emphasize unexpected differential item functioning in two groups after the groups have been matched with respect to the ability or attribute that the item intends to measure (and in any other attribute if necessary). The use of "unexpected" places the focus of attention on comparable examinees among whom differences in performance are not expected. Both approaches use equal ability as measured by total test performance as an operational definition of comparable.

The Standardization Method Applied to the Scholastic Aptitude Test (SAT)

The following notations and terminology are used in the dif methodology. There is always the comparison between two groups of examinees and the performance of one of them will be of primary interest; this will be called the focal group (F) and usually represents the minority group. The performance of the other group, the reference group (R), is considered as the standard against which the performance of the focal group will be compared and it usually represents the majority group.

Basic to all approaches to the study of dif is the idea that only comparable members of F and R should be compared in attempting to identify items that exhibit dif. Only then can it be said that the found differences in the groups are not attributable to differences in the trait. The theory of dif is also based on the belief that groups can be matched at each level of ability on the trait measured by the test. Operationally, comparable levels are defined as identical total test scores in both groups. The performance of each group in the item under study is compared at each score level; the percentage from each group that answered correctly the item at each ability level is the basis for the study of dif.

The standardization index proposed by Dorans and Kulick (1983) is

$$RMWSD = \left[\sum_{s=1}^{S} K_s (P_{fs} - \hat{P}_{fs})^2 / \sum K_s\right]^{1/2}$$

where

RMWSD means the root mean weighted squared differences, since the formula provides the square root of the weighted mean differences in performance at level of ability.

P_{fs} is the proportion of examinees in the focal group at each score level (s) answering the item correctly or the ratio of correct answers to the number of applicants from the focal group at each level of ability as represented

by the total score on the test.

$\hat{P}_{fs}$ is the expected proportion of success in the focal group at each score level (s). This expected proportion is estimated to be equal to the observed proportion in the base group; this means that $\hat{P}_{fs} = P_{bs}$

K_s is the weighting factor represented, generally but not necessarilly, by the number of examinees (K) in the focal group at each level of ability or at each score level (s).

The cutoff value for TMWSD suggested by ETS researchers to consider the index as indicating significant differential item functioning is RMWSD $\geq$ 0.08 and any item providing this value is flagged for detailed consideration as petentially biased.

Rogers & Kulick (1987) present the results of using the standardization method to detect differential item functioning between whites and blacks on the SAT both verbal and math parts and for forms 1, 2 and 3 that were administered during November of 1980, 1983, and 1984. A summary of the data is reported in table 1,

---

Table 1 about here

---

where the items are analyzed both in terms of the standardization index values and also in terms of the absolute differences in item difficulty for the two groups compared.

Results of the Standardization Procedure

Very few items were detected by the standardization procedure as functioning differently in the two groups under consideration. Only about 6.2% of all the items in both the verbal and the math part and in the three forms showed an standardization index equal to or greater than 0.08. This result seems to indicate that, whatever the SAT is measuring, the great majority of the items function in a similar way in both groups of examinimees, at least as dif is defined by the standardization index.

None of the items where the difference in difficulty between the base and the focal groups was 5% or less were detected by the standardization procedure as functioning differently in the two groups. Being the average overall difference in difficulty between the two groups between 13 and 16 percentage points depending on the form and the subtest, it seems that only 5% difference in difficulty in a particular item indicates an aberrant and anomalous departure from the overall pattern of the two groups on the test as a whole, but this anomaly was not detected by the standardization index.

On the other extreme in the scale of difficulty, there were items with differences in difficulty between the focal and the base groups that were 25 percentage points or larger but they were not detected by the standardization index as functioning differently. As a matter of fact, only 20% of the items with group differences in difficulty of 25 percentage points or more were detected as functioning differently.

There is a tendency in the standardization method to detect as functioning differently a greater proportion of items where the difference in difficulty between the focal and the base group is large in relation to the average pass difference than among the items where the difference in pass ratios between the two groups is smaller than the average pass differences. Among the items where the difference in difficulty between the focal and the base groups was larger than 17 percentage points about 10.38% of the items (19 out of of total of 163 items) were detected as functioning differently. Among the items where the difference in difficulty was 10 percentage points or less, only 3.17% (4 out of 126 items) of the items were detected as functioning differently.

A substantial number of items where the group differences in difficulty showed small departures from the average item difference in group difficulty were detected as functioning differently. About 5.52% of the items (12 out of 217) where the difference in difficulty between the base and the focal groups was between 10 and 17 percentage points were detected by the standardization procedure as functioning

differnetly in both groups.

Discussion of the Results

Differential item functioning in two groups as defined by the standardization procedure has nothing to do with differences in item difficulty between the groups. An item may show no difference in the ratios from each group passing the item if the groups are identical in ability level and the item may be detected as functioning differently in the two groups. The difference in the ratios passing the item may be identical to the average group difference in ability and the item may be detected as functioning differently in the two groups. It seems somehow incomprehensible that an item with difficulty levels identical in two groups of the same ability be considered as aberrant or anomalous. It is also difficult to accept that item pass ratios differences equal to the average difference in ability in the groups, as measured by the average total score, be considered also as aberrant or functioning differently. The best available estimate of the ability level of two groups is the average group score; it is reasonably expected that the difference in ability level between two groups will be shown, with small departures, in the groups performance throughout all the items. If the groups are identical in ability, no differences are reasonably expected to exist in the groups performance on each item. If the groups are different in ability, it is reasonable to expect that these group differences will be showm in the performance on each item.

The standardization method compares percentage passing the item at each ability level when the ability is measured by the total score on the test and this difference is represented by $(P_f - P_b)^2$. Psychological measurement instruments are not so precise as to accurately differentiate between consecutive units of the attribute. In a test of moderate length two consecutive scores may indicate the same level in the attribute and the differences between two consecutive scores may be the result of errors of measurement. These errors of measurement are sometimes

positive and sometimes negative and they cancel each other in a total or average passing index but they are added in the standardization index since these differences at each ability level are squared. The standardization index may be an accumulation of errors of measurement instead of an index of aberrant of differential item functioning.

In normal circunstances, the standardization method provides an index of the departure of the difference in difficulty for an item between two groups from the average difference in difficulty for all the items or from the average difference in ability between the groups. The method detects as functioning differently items showing a fixed percetnage point of difference in item difficulty beyond the average item difficulty for the groups. This fixed percentage, in normal circunstances is represented by the 8% proposed by ETS researchers as the value for the RMWSD to be significant. If the difference between the focal and the base groups in percetnage passing the item at each ability level is the same throughout all levels of ability, then $(P_f - P_b)$ becomes a constant in the formula for the standardization index and the index becomes

$$TMWSD = \left[(P_f - P_b)^2 \sum K_s / \sum K_s\right]^{1/2} = (P_f - P_b)$$

and any difference of 8% (0,08) would indicate that the item is functioning differently. When the item is functioning equally throughout all the ability levels, if equally is defined as the same difference between the groups in percentage passing the item at each level of ability, then differential item functioning will be defined for all items and for all groups as a difference of 8% in percentage passing the item.

The Simple Percentage Point Method of Aberrant Item Functioning

The standardization procedure of detecting differential item functioning, regardless of its methodological attractiveness, presents technical problems that

make it unacceptable for detecting
ardization index are: (1) it does
aberrant group differences; (2) it
in passing the item equal to the a
groups; (3) in normal circunstance
will be considered as dif if the di
is 8% beyond the difference in abi
percentage in each group answering
with the standardization procedure
tions dealing with licensing and j
mate of the ability of the groups
assumption will require: (1)substa
both groups in order to obtain sta
general samples in the thousands;
the test is multidimensional inclu
would not indicate a particular ab
indicate the same ability; (3)ther
on the results of the comparison,
total score on the test is used to
test is biased, the procedure will
underestimate the ability level of
are not made between supposely equ

Bias was originally studied
seemed to be unreasonably large gi
of the two groups. Item or test bi
favoring unduly a particular group
ture of bias in testing due to the
adopted methods to eliminate or re
scores (Rooney, 1987). Differencia

biased items. The main problems with the stand-
not detect as dif all the items with large or
considers as dif items with averages differences
verage difference in ability between the two
s, any item, regardless of the groups compared,
fference in passing the item at each ability level
lity level of the two groups as measured by the
correctly all the items. An additional problem
is that it is impractical in most testing situa-
ob selection. The procedure is based on the esti-
at each level of the ability continuum. This
ntial number of examinees at each ability from
ble statistical data and it would translate into
(2) in most licensing and job selection tests
ding more than one factor and a total score
ility and two identical scores do not necessarily
e is a redundancy, without any idea of its impact
when a test is analyzed for possible bias and the
estimate the ability level for each group; if the
overestimate the ability level of one group and
the other group, and the comparisons, therefore,
al levels of ability.

to explain differences in group performance that
ven the sociological and educational charateristics
as, regardless of how it is defined, indicates
. Federal courts have been brought into the pic-
fact that test developers have not voluntarily
duce unnecessary racial differences in test
l item functioning may be an interesting concept

but it does not address the problem of item bias. An item may be functioning differently in two groups of identical ability level even if the percentage of persons passing the item in each group is the same. Consecutive scores do not necessarily indicate different levels of ability and errors of measurement play an unknown role in the scores and the value of the standardization index may be an accumulation of errors of measurement.

In Support of a Simple Method of Detecting Item Bias

Any statistical procedure to determine item bias, in order to be accepted, must make provisions or adjustments for true group differences in the attribute under consideration. Mere group differences in average passing an item, as previously stated, do not constitute sufficient basis for a claim of bias in the item. Statistical elegance, however, with the exclusion of any other consideration have very often been the determining factor in the selection or rejection of a measurement procedure. Methodology can be elevated from handmaiden to the master of the process of inquiry (Nisbert, 1976). Lawshe (1969) found that technically more elegant methods were not better than simpler ones in estimating the reliability, the internal consistency, and the best weihting combination of scores and ended the analysis with a warning: "the statistical nicety of the present methods suggests or implies an order of precision which is not basically inherent in the data; psychological measurements, at this point in time, are quite unreliable; to suggest otherwise by using unwarranted degrees of statistical precision is for the psychologist to delude others and, perhaps, to delude himself (p.122)".

The Golden Rule agreement, if it is interpreted as a court case setting the principles by which items should be selected to be included in a licensure tets but without forcing the particulars of the case on every single test or on every single testing situation, can be defended psychometrically (Gonzalez-Tamayo, 1987). The two principles included in the agreement can be identified as (1) the items included in the test used for a licensure examination should not exceed a certain

level of difficulty to be determined in each testing situation and (2) an aberrant difference in item difficulty between two groups, a difference in percentage passing the item, is an acceptable method of removing items from the test regardless of any other psychometric consideration. It is a well established principle in psychometrics that items discriminate best when their difficulty levels are at the level of ability at which differentiation between test takers is wanted. This ideal difficulty level for a licensure examination should be at the pass mark level and items departing from this level should not be in the test in the first place and if they are they should be eliminated as contributing very little to the discrimination between those who pass and those who fail the test.

The principle that straight differences in passing rates, not any difference but aberrant differences to be determined in every case, to consider an item as bias makes many psychometritians very unconfortable. The most outstanding issue of item bias tehcniques is, however, the effect of removing the bias items on test charateristics; specifically, whether or not the removal reduces the adverse impact of the test on the groups compared without impairing the test predictive validity (Ironson, 1982). Unfortunately, statistical procedures for detecting biased items have become an end in themselves where the results are tabulated and reported and nothing further is done (Scheuneman, 1982). Nungester (1977) reported that removal of items with unacceptable discrimination or difficulty indices, difficulty measured as simple percentage passing the item, improved the reliability and validity of the tests to a greater extent than the removal of items detected as biased by other three psychometric methods. Gonzalez-Tamayo (1984) reported that (1) it is possible to reduce the mean group difference without lowering the validity of the test by eliminating items detected as biased by the simple method of percentage passing the item and (2) a simple criterion of mean group difference in difficulty plus one half of one standard deviation to determine an item as biased worked as well as an a priori criterion in detecting biased items. As stated by Thorndike (1982): "... if the finding of an average difference in item difficulty cannot be used as

the basis for inferring bias, what can be so used? The answer that is often given, and that seems generally reasonable, is that a marked deviation in the discrepancy for a specific item from the average discrepancy for all items may provide such evidence (p.232)", The problem is to define discrepancy or aberrant departure.

A Simple Criterion of Discrepancy

Gonzalez-Tamayo (1984) used the existence or non-existence in the Spanish language of the word representing the item as the criterion for deciding whether the item was biased against Hispanic examinees. The purpose of the study was to develop an item content criterion of item bias and analyze the consequences in the test characteristics of removing such items. The test was the Vocabulary subtest of the California Achievement Test (CAT, 1977). If the word was similar in Spanish and in English the item was judged fair for both groups; if the word appeared only in English it was considered as biased against the minority group. There was the assumption that both groups had had similar opportunity to learn the word if it appeared in both languages, while the words that appeared only in English would be more related to the majority than the minority experience. Schmitt (1987) came to the same conclusion through a reverse procedure. She used the standadization procedure (Dorans & Kulick, 1983) to identify items from the SAT that were functioning differently among Puerto Rican and white candidates. The identified items served as the basis for a hypothesis, that was confirmed, that cognates or word with a common root in English and in Spanish will tend to favor Hispanis examinees item performance.

The Gonzalez-Tamayo (1984) study determined that in the items considered as bias by the external criterion of item content the difference in item passing rate for both groups was about one half standard deviation above the average difference in difficulty between the two groups in the total test. An item in order to be biased must be an outsider, show aberrant or anomalous behavior in relation to the other items (Angoff, 1982). The problem has been to determine the best statistical technique to detect bias. The courts and state legislatures in the last few years

have been requiring that test developers take specific, verifiable, practical steps to reduce unnecessary racial differences in test scores. The simple percentage method does not consider any group difference in passing rates as indicating bias and, on the other hand, it provides a simple, practical and effective way of reducing group differences in test scores without lowering the validity of the test.

An Example of the Simple Percentage Method of Detecting Items as Biased

In an actual examination for a license, it was discovered that tha short answer questions part of the test had an adverse impact on blacks and Hispanics. About 87% of the white applicants againts only 56% of the blacks and 64% of the Hispanics achieved minimum scores in the short answer part of the examination. An item analysis of the pass rates for the three groups revealed extreme degrees of variation from item to item in the percentage of applicants from each group who passed the items. It was arbitrarily defined as aberrant or anomalous difference any difference in percentage passing the item that was greater than 1.5 SD beyond the average mean. This represented 27 percentage points difference in passing the item when the average difference was only 10.76 percentage points. A total of 10 items were eliminated from the test following this procedure, out of a total of 130 item in the test. The eliminated items showed a point biserial correlation in the upper 20s (2 items), upper 30s (3 items) and in the 40s (5 items) indicating that high discrimination and aberrant item performance go together when the groups possess different amount of ability as measured by the total score on the test analyzed for possible bias,

Table 2 contains the data about the number of applicants from each of the two groups analyzed who achieved minimum scores in the total test (130 items) and in the reduced test (120 Items) after eliminating the ten items detected as functioning anomalously. The result of eliminating the ten items was that new applicants who

had previously passed the test failed to achieve minimum scores after the elimination of the biased items; there were also applicants achieving minimum scores who originally had failed to pass the test. The net effect of eliminating the ten items was negative for the white group and positive for the other groups. Eleven applicants, eight of them white, passed the written test because of ten items that were functioning anomalously in relation to other items. Twenty two applicants, at least 12 of them from the protected groups, failed to achieve minimum scores because of the ten items that were functioning differently. It is difficult to defend that 7.7% of the items in the test, that seem to be outsiders in their functioning, be allowed to have such an important and disproportionate impact in the outdome of who reaches or fails to reach a minimum score.

The pass rate ratios, or the ratios between the pass rates for the protected groups and the white group , improved substantially with the elimination of the items functioning anomalously. The improvement in the pass rate ratio for blacks was from 0.64 to 0.73 and for Hispanics the improvement was from 0,73 to 0,84. It is important to note that the elimination of biased items did not eliminate the group differences in mean score nor the pass rate differences but merely reduced them ro more reasonable levels. In conclusion, the application of a direct, simple method of detecting biased items was effective in reducing the adverse impact of the test on the protected groups without eliminating group differences.

Table 1: Number of Items from the SAT(Three Forms) Verbal and Math Parts at Various Levels of Difficulty Differences between Whites and Blacks and at Different Values of the Standardization Index

Group Differences in Item Difficulty $P_b - P_f$	Total Number of Items	Standardization Index ≥ 0.08
5% or less	40	0
5.1% to 10%	86	4 (4.65%)
10.1% t0 17%	146	4 (2.74%)
17.1% to 20%	71	8 (11.27%)
20.1% to 25%	57	4 (7.02%)
over 25%	35	7 (20%)
TOTAL	435	27 (6.21%)

Table 2: Number of Applicants Passing the Total Test and the Reduced Test After Eliminating the Items Functioning Anomalously

	Whites	Blacks	Hispanics	No Inform.	Total
Total Test (130 items)					
Total Applicants	343	93	80	140	656
Number passing	300	52	51	101	504
Pass Rate	0.87	0.56	0.64	0.72	0.77
Pass Rate Ratio	1.00	0.64	0.73		
Reduced Test (120 items)					
New Failures	8	1	1	1	11
New Passers	3	6	6	7	22
Net Change	-5	+5	+5	+6	+11
Number Passing	295	57	56	107	515
Pass Rate	0.83	0.61	0.70	0.76	0.79
Pass Rate Ratio	1.00	0.73	0.84		

BIBLIOGRAPHY

Anastasi, A. 1982. Psychological Testing (5th Edition). McMillan Publishing Co, New York.

Angoff, W.H. 1982. Use of Difficulty and Discrimination Indices for Detecting Item Bias. In R.A. Berk (ed.) Handbook of Methods for Detecting Test Bias. The John Hopkins University Press, Baltimore, Maryland.

Berk, R.A. (Ed.), 1982. Handbook of Methods for Detecting Test Bias. The John Hopkins University Press, Baltimore, Maryland.

California Achievement Test. 1977. CTB/McGraw Hill, California.

Dorans, N.J & Kulick, E. 1983. Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977. An Application of the Standardization Approach. Research Report N. RR 83-9. Princeton, New Jersey: Educational Testing Service.

Dorans, N.J. & Kulick, E. 1986. Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Gonzalez-Tamayo, E. 1984. An Item Content Criterion of Debiasing a Test. ERIC Document Reproduction Service. No. Ed. 247-240.

Gonzalez-Tamayo, E. 1987. The Golden Rule Agreement is Psychometrically Defensible. ERIC Document Reproduction Service. No. Ed. 283-868.

Holland, P.W. & Thayer, D.T. 1988. Differential Item Performance and The Mantel Haenszel Procedure. In H.Wainer and H.I. Braun, Test Validity. Lawrence Erlbaum Ass., Hillsdale, New Jersey.

Ironson, G. H. 1982. Use of Chisquare and Latent Trait Approaches for Detecting Item Bias. In R.A. Berk (Ed.) Handbook of Methods for Detecting Test Bias. The John Hopkins University Press, Baltimore, Maryland.

Jensen, A. R. 1980. Bias In Mental Testing. The Free Press, New York.

Jensen, A.R. 1982. Test Bias: Concepts and Criticisms. In C.R. Reynolds & R.T. Brown (Eds.) Perspectives of Bias in Mental Testing. Plenum Pres: New York,

Lawshe, C. H. 1969. Statistical Theory and Practice in Applied Psychology. Journal of Personnel Psychology, 22, 117-124.

Mantel, N. & Haenszel, W. M. 1959. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of National Cancer Institute, 22, 719-748.

Nisbert, R. 1976. Sociology as an Art Form. Oxford University Press, New York.

Nungester, R.J. 1977. An Empirical Examination of Three Methods of Item Bias. Unpublished Doctoral Dissertation, Florida State University.

Reynolds, C. R. 1982. The Problem of Bias in Psychological Assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), The Handbook of School Psychology. Wiley, New York.

Rogers, H.J. & Kulick, E. 1987. An Investigation of Unexpected Differences in Item Performance between Blacks and Whites Taking the SAT. In A.P. Schmitt & N.J. Dorans (Eds.), Differential Item Functioning on the Scholastic Aptitude Test. ETS, R-M 87-1, Priceton, New Jersey.

Rooney, P.J. 1987. A Response from Golden Rule to "ETS on Golden Rule". Educational Measurement, Issues and Practice, 6, 19-23.

Scheuneman, J.D. 1982. A Posteriori Analysis of Biased Items. In R.A. Berk (Ed.), Handbook of Methods of Detecting Test Bias. The John Hopkins University Press, Baltimore, Maryland.

Schmitt, A. P. 1987. Unexpected Differential Item Performance of Hispanic Examinees. In A.P. Schmitt & N.J. Dorans (Eds) Differential Item Functioning on the Scholastic Aptitude Test. ETS, R-M 87-1, Priceton, New Jerse.

Scriven, M. 1980. Method of Inquiry in the Philosophy of Education. In R.M. Jaeger (Ed,) Alternative Methodologies in Educational Research. Washington D.C.: American Educational Research Cassette Service.

Thorndike, R.L. 1982. Applied Psychometrics. Houghton Mifflin, Co, Boston.